Research and Prediction on China's Novel Coronavirus (2019-nCoV/ COVID-19) Epidemic——Based on Time Series ARIMA Model

Wenbohao Zhu[1, a], Xiaofeng Li[1, b] and Bo Sun[2, c*]

[1]School of Economics and Trade, Guangdong University of Foreign Studies, Guangzhou, China

[2]International Service Outsourcing Research Institute, Guangdong University of Foreign Studies, Guangzhou, China

[a]13032264827@163.com, [b]phdlxf@126.com, [c]sbgz168@gdufs.edu.com

[*]corresponding author

Keywords: ARIMA Model; China's Novel Coronavirus (2019-nCoV) Epidemic; Prediction

Abstract: In order to analyse and predict the short-term trend of China's 2019 Novel Coronavirus (2019-nCoV/ COVID-19) epidemic, the study uses historical statistical data of China's COVID-19 cases between December 1, 2019 and July 31, 2020. Based on the characteristics and trends of its time series, a differential integrated moving autoregressive average model (ARIMA) is established and used with STATA 16.0. This model detects and predicts the future short-term data of China's COVID-19 epidemic situation. In the first half of Augusts, the maximum number of newly confirmed cases in China should be 319 and the minimum should be 199. There will be no erratic fluctuation or sudden large increases in the number of new infected individuals or accumulative confirmed cases. The research proves that, the prevention and control measures of China's COVID-19 epidemic by the Chinese government are effective, epidemic has been gradually controlled and defence against disease has entered normalization. The analysis and prediction on China's COVID-19 epidemic based on the ARIMA model can help China better respond to the outbreak of COVID-19 epidemic in the short term and provides decision-making suggestions for control of the disease in short term.

1. Introduction

The 2019 Novel Coronavirus (2019-nCoV/ COVID-19) Epidemic (hereinafter referred to as " COVID-19 epidemic" or "epidemic") was first detected in Wuhan, Hubei Province in December 2019. In mid and late January of2020, the epidemic was leaping rapidly, and spread to other parts of China in large scale (Figs 1 and 2). Emergency measures were deployed as the level of prevention and control has been continuously improved. By 24:00 on July 31 2020, China has reported 88077 confirmed cases, 4668 cases of death and 81213 cases of rehabilitation. The epidemic is a major public health emergency (Level II) with the widest range of infection (Figure 1), the fastest transmission speed (Figure 2), and the most difficult to cure and prevent (Figure 3) since the establishment of Chinese government.

So far, although the COVID-19 epidemic situation has been basically controlled under the joint efforts and cooperation of the whole country, due to the lack of sufficient prediction and judgment in the initial stage of the outbreak, the COVID-19 epidemic has had tremendous social and economic impact on China, and also brought heavy disasters to the people of the world. Therefore, it is necessary to establish a good prediction model based on the time series characteristics of COVID-19 epidemic and its development trend, so as to monitor and warn the basic situation of the outbreak in the near future as early as possible, and provide a scientific basis for the follow-up epidemic prevention and control, and stabilize the phased achievements of the epidemic.

The first part is the research background, purpose and significance. The second part is literature review, which mainly introduces ARIMA model and its application. It also lists the research and application of ARIMA model in the field of major disease prediction. The third part is data analysis,

which details the sub item comparative data of COVID-19 epidemic in China in the initial period, high incidence period and the latest period. The fourth part is ARIMA modeling process, including ARIMA model parameter estimation method, prediction and residual test process. The fifth part is, based on the practical data of ARIMA model prediction with Stata, the final conclusion is that the model and the actual value fit well. The results pass the significance and validity test, and at the same time, the forecast value is obtained. The sixth part is, based on the ARIMA model fitting value of the research conclusion, the paper puts forward suggestions for epidemic prevention and control. In the end, this paper points out the limitations of the study and what can be improved.

The innovation of this study is, to analyze the historical data of the cumulative number of confirmed cases and the newly diagnosed cases in China, according to the characteristics of its time series data, the ARIMA model was established for short-term prediction of the COVID-19 epidemic situation, and the daily mortality rate was calculated on this basis, so as to strengthen the prevention and control measures and monitor the new epidemic situation in the later stage by using the new data.

2. Literature Review

2.1 ARIMA Model

ARIMA model refers to auto regressive integrated moving average, which benefited from contributions of research on autoregressive (AR)model by British statistician G.U.Yule and research on moving average (MA) model by British mathematician G.T.Walker, the core idea of regression analysis and moving average in AR and MA model are the foundation of ARIMA model. On the basis of these studies, the statisticians G.E.P.Box and G.M.Jenkins jointly proposed ARIMA model, its application principle and method in 1970, thus ARIMA is also known as the Box-Jenkins model. Due to the limitation of ARIMA model in the processing of time series, statisticians and econometrics relaxed the conditional assumptions of ARMA model (including ARIMA), such as univariate and homo-variance, and successively put forward to ARCH model, GARCH model (heteroscedasticity situation), cointegration theory (multivariate situation) and threshold autoregressive model (for nonlinear time series) with more relaxed conditional assumptions of time series, these models have become the classic contents of time series.

2.2 The Application of ARIMA Model

ARIMA model is widely used in economic, financial, environmental protection, transportation, medicine, aviation and other fields of time series prediction. This model is widely used as it is relatively simple and easy to obtain the required data with high prediction accuracy.[1]. Based on the data of China's iron ore price index (CIOPI), Yang et al. established ARIMA (3,1,3) model to study and forecast the fluctuation and trend of China's iron ore price index in the near future, the results show that the predicted value and the actual value fit well, and the average relative error is 1.09% [2]. Zhang et al. used SPSS software to study the time series of death rate in traffic accidents in China, the results show that ARIMA model is accurate and can provide scientific prevention basis for reducing the death rate of traffic accidents [3]. Based on the London spot gold price in the historical period, Xu et al. established ARIMA model to predict the gold price trend in the first half of 2011, providing an effective basis for China's gold reserve policy decision-making [4]. Song et al. used ARIMA model combined with BP neural network theory to predict the indexes of sulfur dioxide, nitrogen dioxide and inhalable particles of air quality in Baotou city, the results showed that the short-term prediction effect was good, which could effectively provide scientific basis for the prediction of air pollutants [5]. Shi et al. used the CRITIC method to combine the long short-term memory (LSTM) with ARIMA model to improve the prediction accuracy of short-term flight path [6].

2.3 ARIMA Model in the Field of Major Disease Prediction

Based on the number of new HIV / AIDS cases reported by China Center for Disease Control and prevention, Yin et al. used ARIMA (0,0,0) (0,1,0) (12) and ARIMA (0,1,1) (0,1,0) (12) to predict the routes of heterosexual and homosexual transmission of HIV / AIDS [7]. Zhou et al. used the exponential smoothing model and ARIMA model to study and predict the infection rate of

human and livestock Schistosoma in Hunan Province, the results showed that the two models had good fitting effect [8], which further verifies the effectiveness of ARIMA model in the prediction of major diseases and epidemic situation. In order to analyze the incidence of pneumoconiosis in Jiangsu Province, Bian et al. established the ARIMA-GM model based on the prediction residuals of ARIMA model to analyze and study the incidence of pneumoconiosis, which is the most harmful occupational disease in China [9]. Ma et al. established the ARIMA (2,1,1) (0,1,1) (12) model to predict incidence rate of syphilis with good fitting results [10]. He et al. established ARIMA model based on the statistical data of rabies incidence of farmers, students and scattered-residing children in China from 2004 to 2013, and predicted the number of rabies cases from 2015 to 2017 based on the data in 2014 as validation [11]. Ding et al. studied the data of measles in recent twenty years, and predicted the incidence rate of measles after the measles vaccine was intensified using ARIMA (0,0,0) (0,1,0) model [12].

3. Data Analysis and Modeling Steps

3.1 Data Analysis

In order to further analyze the epidemic situation in 2020 and its future trend, the data obtained from the Chinese National Health Committee from December 2019 to July 2020, are sorted out including the number of daily cumulative diagnoses, deaths and rehabilitation. According to the fluctuation characteristics of the daily new death toll, the initial period of the epidemic situation is from December 1, 2019 to January 15, 2020, the epidemic period is from January 16, 2020 to February 29, 2020, and the latest epidemic situation is from June 15, 2020 to July 31, 2020.

$$\text{case fatality rate} = \frac{\text{the number of deaths on the day}}{\text{the number of deaths on the day} + \text{the number of new rehabilitation}} \quad (1)$$

Through calculation, we find that the number of newly diagnosed cases per day and the mortality rate of the day（Eq.1）have been significantly reduced compared with the high incidence time (middle and late January) in the latest epidemic (middle and late July). In July, the mortality rate was significantly 0 in 20 days, the highest value was 0.2857, and the lowest value was 0.0128. Compared with the highest value of 1 and the lowest 0.1667 (except 0) on January, the mortality rate has decreased significantly. In July, the highest value of newly diagnosed cases was 109 cases, and the lowest value was 0 cases; compared with the highest value of 1933 cases and the lowest 0 cases in January, it has decreased significantly. This further proves that the prevention and control measures for the epidemic in China are effective, and the epidemic situation has been gradually controlled.

3.2 Model Principle

ARIMA model is mainly effective for stationary time series, because the idea of autoregression is based on the fact that the current value of time series is highly dependent on the historical value in the past, while the moving average algorithm just eliminates the prediction fluctuation, which effectively handles the prediction error of autoregression. Therefore, we need to determine the order of ARMA model through a series of screening processes. After selecting the appropriate model, we can compare it with historical data to see whether it fits or not, then test the residual error of the predicted value, and finally get an effective model with good fitting.

3.3 Modeling Steps

Stationarity test: visualize the original data and judge the stability of the sequence according to the sequence diagram and square root test. When the time series is periodic or has a significant trend, the unit root test is carried out directly; otherwise, the first-order difference is carried out on the original sequence data, and if necessary, the second-order or multi-level (determine d) is used to determine the order and stabilize it;2. Fitting ARMA model: according to the properties of autocorrelation graph (ACF) and partial autocorrelation (PACF), ARMA (p, q) model with appropriate order is selected for fitting (in this study, model is estimated under MLE principle);3. Using the model, the predicted value is established, and the fluctuation and trend of the predicted

value are observed and compared with the historical value;4. Test the residual sequence of predicted values, including validity test and stability test5. Make a new short-term forecast for the future;6. Modeling completed.

4. Model Screening and Prediction

4.1 Stationarity

We take the daily newly added diagnosis number as the variable, through observing the trend of its time series, we preliminarily judge that it is a stationary sequence (there is no upward or downward trend, and there is obvious volatility). In order to further examine the stabilization of the series, we carry out unit root test on the logarithm of the original data. The Z-test of the new confirmed number sequence is less than the critical value of each test, and the P-value is less than 0.05. In the same way, the sequence of cumulative number of confirmed patients is not stable, so the first-order difference and the unit root test are carried out. (the result is the original sequence of the newly diagnosed number).

Table 1. Unit root test after first-order difference of cumulative number of confirmed cases

	Test Statistic	1% Critical Value	5% Critical Value	10%CriticalValue
Z(t)	-7.985	-3.464	-2.881	-2.571

MacKinnon approximate p-value for Z(t) = 0.0000

4.2 Determinable Coefficient

The unit root test is used to test the data after the first-order difference of the cumulative number of confirmed patients and time series has become stationary as seen in the Table 1. In the model ARMA (p, d, q), d is the difference times when the time series becomes stationary, so d = 1 in this study. Taking the cumulative number of confirmed patients as the main time series, the predicted value of the newly diagnosed number can be directly obtained. On this basis, the predicted value of the cumulative number of confirmed patients in the future period can be obtained by selecting the appropriate base period. The first-order lag autocorrelation graph and partial autocorrelation graph of the cumulative number of confirmed patients are investigated and both ACF and PACF are tailed. We use ARMA model.

By using the information criterion (AIC and BIC) to screen the lag order (Table 2), it can be seen that AIC and BIC in ARIMA (1,1) are the minimum (optimal), and the comprehensive conclusion is that p = 1, q = 1, d = 1 are the optimal parameters of ARIMA model. The coefficient of AR is 0.944 and the coefficient of MA is -0.662, which is highly significant.

Table 2. Screen the lag order to select P and Q with information criterion

	(1) Arma10	(2) Arma11	(3) Arma20	(4) Arma21	(5) Arma30	(6) Arma31	(7) Arma40	(8) Arma41
d_confirmed	353.0	325.4	348.6	324.9	342.9	325.6	337.6	325.9
cons	(0.61)	(0.19)	(0.35)	(0.19)	(0.26)	(0.19)	(0.23)	(0.19)
ARMA								
L.ar	0.574***	**0.944*** **	0.403***	0.947***	0.339***	0.899***	0.309***	0.841***
	(21.81)	**(26.47)**	(12.77)	(11.62)	(9.77)	(7.97)	(8.56)	(4.63)
L2.ar			0.296***	-0.00209	0.210***	-0.0240	0.180***	-0.00696
			(11.32)	(-0.04)	(6.82)	(-0.44)	(5.48)	(-0.10)
L3.ar					0.213***	0.0618	0.165***	0.0449
					(8.08)	(1.34)	(5.58)	(0.92)
L4.ar							0.139***	0.0483
							(4.05)	(0.85)
L.ma		-0.662***		-0.663***		-0.615***		-0.559**

		(-12.91)		(-8.55)		(-5.37)		(-3.03)
sigma	979.4***	896.6***	934.7***	896.6***	912.8***	895.5***	903.7***	894.8***
cons	(62.02)	(73.31)	(58.33)	(69.21)	(63.01)	(67.82)	(68.90)	(67.49)
N	250	250	250	250	250	250	250	250
AIC	4159.3	**4117.6**	4138.2	4119.6	4128.5	4121.0	4125.5	4122.6
BIC	4169.9	**4131.7**	4152.3	4137.2	4146.1	4142.1	4146.6	4147.3

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Forecast. Compare the fitting line between the predicted value generated by ARIMA model and the original value: blue line is the fitting line y of the predicted value, and red line is the original value, that is, the new confirmed number. The COVID-19 epidemic data series was predicted by ARIMA (1, 1, 1) model in the next two weeks. According to the comparison between the real value and the fitting value of the daily newly added confirmed number (Figure 1), the predicted data is in good agreement with the actual data.

Significance Test. Carry out unit root test on the residuals of fitting series. According to Z-test and p-value, the residuals are stationary time series, and they are also stationary series according to fitting residual (Figure 2).

Validity Test. using Q statistics to test the white noise, Portmanteau (Q) statistic is 8.6710, and Prob > chi2(40) =1.0000, there is no autocorrelation of the disturbance term, cumulative periodogram white-noise test is also past, so it can be seen that the fitting effect of the model is good.

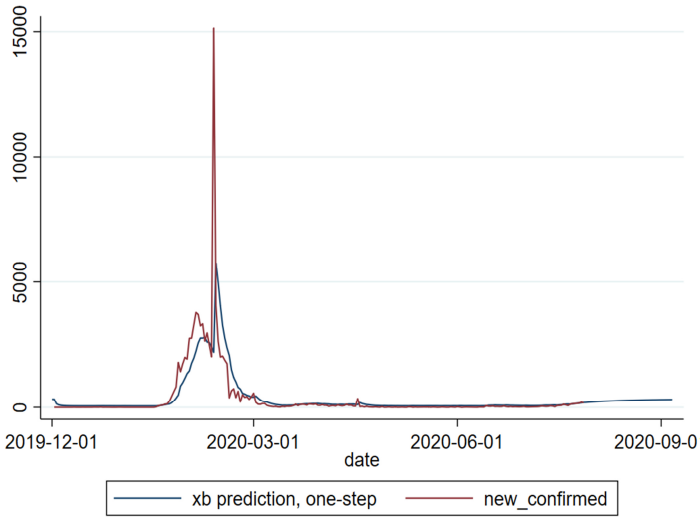

Figure 1. true value and the fitting value of the daily newly added confirmed number

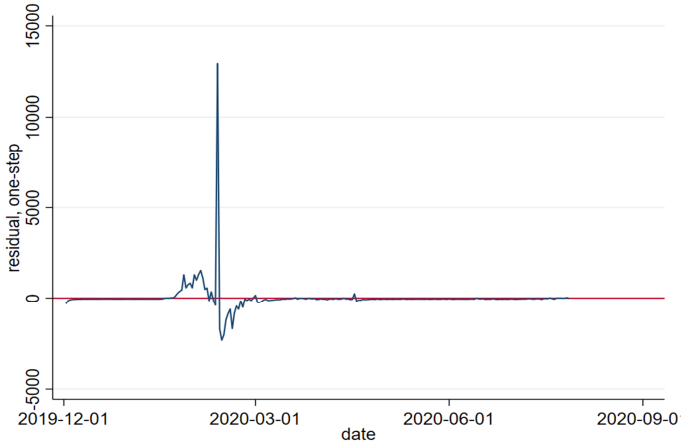

Figure 2. Fitting residual

5. Conclusions and Suggestions

The model ARIMA (1, 1, 1) is obtained and established to predict the COVID-19 epidemic situation in the short term. The forecast results show that in the first half of August, the maximum number of new confirmed cases will be 240 and the minimum is 164. There will be no sudden increase in the number of new cases, which is more consistent with the actual value. The COVID-19 epidemic model ARIMA (1, 1, 1) has a high degree of fit from December 1, 2019 to July 31, 2020. The model ARIMA (1, 1, 1) can be used to detect the short-term fluctuation and future trend of COVID-19 epidemic in China. It can provide reference for the prevention and control of epidemic situation and short-term judgement, and provide a policy basis for the future intervention decision of China's epidemic. At the same time, the above results also show that the prevention and control measures of COVID-19 epidemic in China are effective, and the epidemic situation has been gradually controlled.

However, in view of the positive growth of the follow-up predicted results, it is suggested that, on the one hand, China should consolidate the phased achievements of the current epidemic situation, on the other hand, it should formulate prevention and control measures to strictly prevent imported cases and prevent the second cluster outbreak of the epidemic. ARIMA model has certain feasibility in epidemic prediction, but it is limited by the applicable period, which is more suitable for short term. Therefore, after the epidemic data are constantly updated, it should be revised with new data to adapt to a new round of short-term forecasting and provide preventive advice. In addition, ARIMA-GARCH model or ARIMA model combined with other models can be also used to control the error in present model to revise the parameters.

Acknowledgements

This paper is supported by the 13th five-year plan of philosophy and social sciences of Guangdong Province "Empirical Study and Countermeasures on International Competitiveness of Service Trade based on Guangdong Cultural and Creative Industries" Project No.GD17XYJ14；it is a phased achievement of Guangdong Soft Science Project "Research on Methods and Implementation Path of International Scientific and Technological Cooperation in Guangdong, Hong Kong and Macao Bay Area" in 2020, Project No. 2020A1010020041，a phased achievement of The 2019 Guangzhou Social Science Planning Project "Empirical Research on Guangzhou Science and Technology Finance under the Synergy Mechanism of Scientific and Technological Innovation and Technology Finance" Project No.2019GZGJ59, and a phased achievement of Guangdong Social Science Planning Project "Research on the Constraints and Optimization Path of Knowledge Transfer in Universities in Guangdong, Hong Kong and Macao Bay Area", Project No. GD20CJY04.

References

[1] Walter Enders. Applied Economic Time Series. Technometrics, Vol. 37(1995) No. 4, p.469.

[2] J.W. Yang and Y. Li. Prediction of China's Iron Ore Price Index based on ARIMA model. Practice and Recognition of Mathematics 50.11 (2020): 289-298.

[3] J. Zhang, X.M.Liu，Y.L.He and Y.S.Chen. Application of ARIMA Model in Traffic Accident Prediction. Journal of Beijing University of Technology. 12 (2007): 1295-1299.

[4] L.P. Xu and M.Z. Lu. Short Term Analysis and Prediction of Gold Price based on ARIMA Model. Financial Science. 01 (2011): 26-34.

[5] Y.C. Song and S. Zhen. Appplication of BP Neural Network and Time Series Model in Air Quality Prediction of Baotou city. Resources and Environment in Arid Area. 27.07 (2013): 65-70.

[6] Q.Y. Shi, J.C.Yue, P.Han and W.Q. Wang. Short Term Flight Trajectory Prediction based on

LSTM-ARIMA Model [J]. Signal Processing, 2019,35 (12): 2000-2009.

[7] S.Q.Yin, X.Y.Yan, H.X.Su and B. Zhang. Prediction of New Cases of Sexually Transmitted HIV / AIDS based on ARIMA Model [J]. Chinese Journal of AIDS and STD, 2020,26 (07): 709-713.

[8] J.Zhou et al. Application of Exponential Smoothing and ARIMA model in Prediction of Schistosomiasis Epidemic Trend in Hunan Province [J]. Chinese Journal of Schistosomiasis control, 2020,32 (03): 236-241 + 254.

[9] Z.L. Bian et al. Analysis of Pneumoconiosis in Jiangsu Province by ARIMA-GRNN Model. Environmental and Occupational Medicine. 36.08 (2019): 755-760.

[10] X.M. Ma et al. Prediction on the Incidence Rate of Syphilis Monthly based on ARIMA Model. Journal of Xi'an Jiao Tong University (Medical Science). 39.01 (2018): 131-134+152.

[11] J. He et al. Application of ARIMA Model in Prediction of Rabies Incidence in Different Populations. Chinese Journal of Occupational Diseases of Labor Health 36.07 (2018): 512-515.

[12] X.Y.Ding et al. Prediction and Analysis of Measles Epidemic Using Time Series Model. Journal of Nanjing Medical University (Natural Science Edition) 31.08 (2011): 1200-1203.